Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30608

A Comparative Study on the Performance of BERT Sentiment Classification Models Based on Data Augmentation

Chong Wang

Chongqing University, Chongqing, China
exusiair2@gmail.com

Abstract. With the development of social media and e-commerce platforms, text sentiment analysis is of great value in public opinion monitoring, market analysis, and user experience research. Pre-trained language models such as BERT significantly improve sentiment classification performance, but are still prone to overfitting in small sample scenarios [1]. In order to improve the robustness and generalization ability of the model, this paper compares the performance differences of three lightweight data enhancement methods, namely synonym substitution, random masking and backtranslation, in Chinese sentiment classification tasks [2,3]. Based on the ChnSentiCorp dataset, the enhancement strategy is analyzed from three dimensions: semantic retention, computational cost and performance improvement. The results show that both synonym substitution and random masking can stably improve the model performance, while the effect of unfiltered backtranslation is significantly reduced due to semantic shift. After adding semantic similarity and sentiment consistency screening, the backtranslation performance is restored. This paper provides suggestions for the selection of enhancement strategies for emotion classification in small sample scenarios.

Keywords: Data augmentation, BERT, Sentiment classification, Few-shot learning, Back-translation

1. Introduction

Sentiment analysis, as a key direction in natural language processing, is widely used in public opinion analysis, product reviews, and public sentiment monitoring. In recent years, pre-trained language models based on Transformer architectures (such as BERT) have demonstrated outstanding performance across various NLP tasks [1,4]. However, acquiring high-quality annotated Chinese corpora is costly, often leading to insufficient model generalization in low-resource scenarios.

Data augmentation is considered an effective approach to alleviate the low-resource problem and has been applied in machine translation and text classification [3,5]. Yet most existing research focuses on English, and systematic comparisons of lightweight augmentation methods for Chinese low-resource sentiment classification remain limited. Especially in the Chinese field, there are great differences in the performance of different enhancement methods in terms of semantic preservation, ambiguity introduction, and style transfer, which makes it necessary to evaluate the enhancement strategy in detail.

DOI: 10.54254/2755-2721/2026.TJ30608

This paper focuses on three easily implementable augmentation methods, combined with the BERT fine-tuning framework, to investigate their applicability and differences in low-resource sentiment classification tasks.

2. Methodology

2.1. Overview of BERT

BERT leverages a bidirectional transformer structure for MLM and NSP pre-training, providing rich contextual semantic representations [1]. In the text classification task, the [CLS] vector is connected to the fully connected layer and the full model is fine-tuned. In practical applications, stable training of BERT is particularly important under small sample conditions because small-sample training tends to lead to rapid convergence and overfitting of the model in early epoch, which is also a key background for the role of data augmentation methods.

2.2. Data augmentation strategies

2.2.1. Synonym substitution

Find synonyms based on words or context vectors and generate new samples with similar semantics. This method is simple and efficient, and it is one of the most commonly used data enhancement methods [2].

2.2.2. Random mask

Randomly masking some token in the text and predicting the replacement words by the language model, similar to BERT's MLM task, is helpful to improve the robustness of the model.

2.2.3. Back translation

Translating texts into another language and back to the original language is often used to generate high-quality semantic diversity samples [3,6]. However, if the translation quality is unstable, it may cause semantic drift or even emotional inversion.

3. Experiments

3.1. Dataset

It adopts ChnSentiCorp and contains about 10k Chinese comments. In order to simulate small samples, the training set is reduced to 10 – 20%, and the rest samples are generated by enhancement. The dataset includes hotel reviews, product reviews, and other scenarios, and the text length is evenly distributed, making it suitable as a testing platform for small-sample enhancement strategies.

3.2. Experimental settings

- Pre-training model: BERT - Base - Chinese
  - Hyperparameters: maxlen = 128, batch = 32, lr = 2e-5, epoch = 3.
  - Evaluation indicators: Accuracy, Precision, Recall, F1.

DOI: 10.54254/2755-2721/2026.TJ30608

Four types of experimental groups were set up: baseline, synonym substitution, random mask and back translation.

4. Results and discussion

4.1. Overall performance comparison

Synonym substitution has the best effect, followed by random mask; The performance of unfiltered back translation is seriously degraded due to semantic noise.

Table 1. Performance of the 4 groups

Group	Accuracy	Precision	Recall	F1
Baseline	0.915	0.916	0.915	0.915
Synonym substitution	0.927	0.927	0.927	0.927
Random mask	0.921	0.921	0.921	0.921
Back-translation	0.498	0.248	0.498	0.332

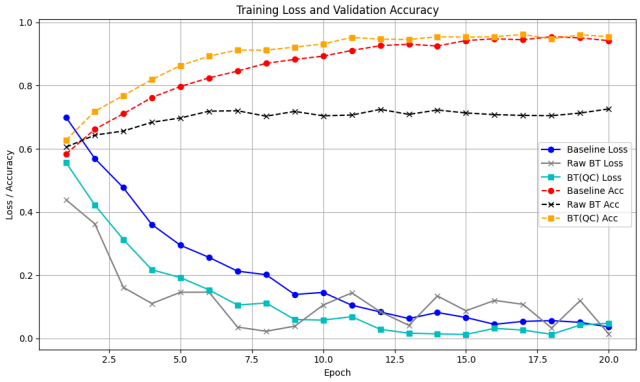

Figure 1. Training loss and validation accuracy

4.2. Error analysis of back-translation samples

The manual and automatic index analysis of 7316 retranslated samples shows that:

- Semantic deviation: 45%
- Emotional inversion: 37%
- Abnormal sentence pattern: 28%

Consistent with the related research in the field of machine translation, unfiltered back translation may produce a large number of invalid or even erroneous samples, leading to model performance collapse [6].

4.3. Improved back-translation with quality filtering

After the semantic similarity (threshold 0.8) and emotional consistency screening mechanism are adopted, the quality of the retranslated samples is significantly improved.

The filtered retranslation results recover and surpass baseline, which verifies the effectiveness of high-quality retranslation.

DOI: 10.54254/2755-2721/2026.TJ30608

Table 2. Re-experiment results

Group	Accuracy	Precision	Recall	F1
Back-translation (unfiltered)	0.498	0.248	0.498	0.332
Back-translation (filtered)	0.922	0.922	0.921	0.921

5. Discussion

The experimental results show that:

• Simple enhancement methods (synonymous substitution, random mask) can stably improve the performance of the model and have low computational cost.

• The effectiveness of back translation is highly dependent on quality control, and it will damage the model performance without screening.

• In the emotional classification of small samples, it is more important to enhance the "quality" of samples than the "quantity".

The results are consistent with the existing enhancement research, that is, the benefits of data enhancement are directly related to the semantic retention of generated data [7].

6. Conclusion

In this paper, three data enhancement strategies are systematically compared in the task of Chinese small sample emotion classification. The results show that synonym substitution and random mask can effectively improve the performance of BERT model. Back translation needs to cooperate with semantic and emotional screening to play its role. The research provides a practical reference for selecting the appropriate lightweight enhancement scheme in the task of text classification with low resources.

References

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186). Association for Computational Linguistics.

[2] Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6382-6388). Association for Computational Linguistics.

[3] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 86-96).

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems 30 (pp. 5998-6008).

[5] Wang, Y., Lu, Y., & Wei, F. (2021). A survey on low-resource neural machine translation. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21) (pp. 4390-4397). International Joint Conferences on Artificial Intelligence.

[6] Edunov, S., Ott, M., Baevski, A., Fan, A., & Ge, M. (2018). Understanding back-translation at scale. arXiv preprint arXiv: 1808.09381.

[7] Wu, W., Zhang, X., Zhao, Z., & Chen, H. (2019). Conditional BERT contextual augmentation. In Proceedings of the 35th International Conference on Computational Science (ICCS) (pp. 526-537). Springer.